7th Statistics on Health Decision Making: Epidemiology

July 10 › 11, 2025 | University of Aveiro

DIRECTED ACYCLIC GRAPHS AS A CAUSAL INFERENCE TOOL:

From principles to applications

Andreia Leite

National Institute of Health Research Doctor Ricardo Jorge

NOVA National School of Public Health

Outline

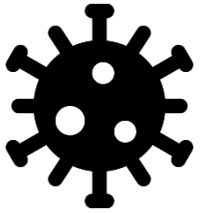
- Causal questions;
- Confounding;
- Traditional approaches;
- Directed acyclic graphs – notation and rules;
- How to build a DAG;
- Use of DAGs;
- Overview.

Causal questions in epidemiology

What is the effect of closing schools on the COVID-19 pandemic control?

What is the effect of the “sugar tax” on the sugared beverages consumption?

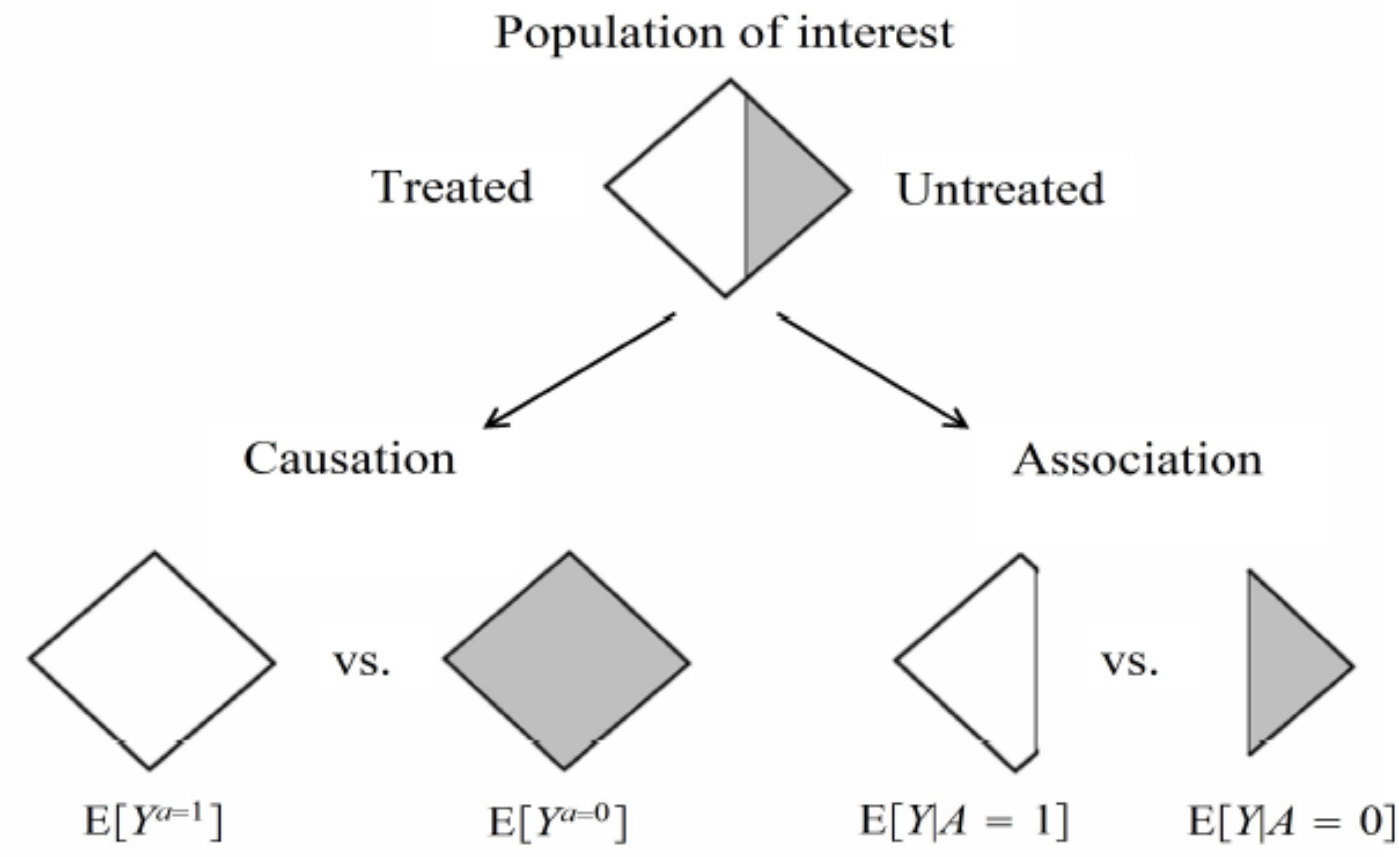
What is the effect of HIV on stroke?



What is the effect of COVID-19 vaccine on COVID-19 admissions?

Virus icon from Flaticon

Causation vs. association



Hernán M, Robins J. Causal inference: What if.

The issue with confounding

- Various meanings in different areas;

- Distorts the relationship between the exposure and the outcome;

- Traditional definition:
  a) it must be associated with the exposure;
  b) it must be associated with the outcome in the unexposed;
  c) it must not lie on a causal pathway between exposure and outcome.

How to identify confounders?

Before applying a statistical correction method, one has to decide which factors are confounders. This sometimes is a complex issue (11-13). **Common strategies to decide whether a variable is a confounder that should be adjusted or not, rely mostly on statistical criteria.** The research strategy should be based on the k l framework an uld be involved for evaluating the confounders. Statistical models (especially regression models) are a flexible way of investigating the separate or joint effects of several risk factors for disease or ill health (14).

menti.com | 7589 8195

Pourhoseingholi MA et al. Gastroenterol Hepatol Bed Bench 2012;5(2):79-83

How has it been handled

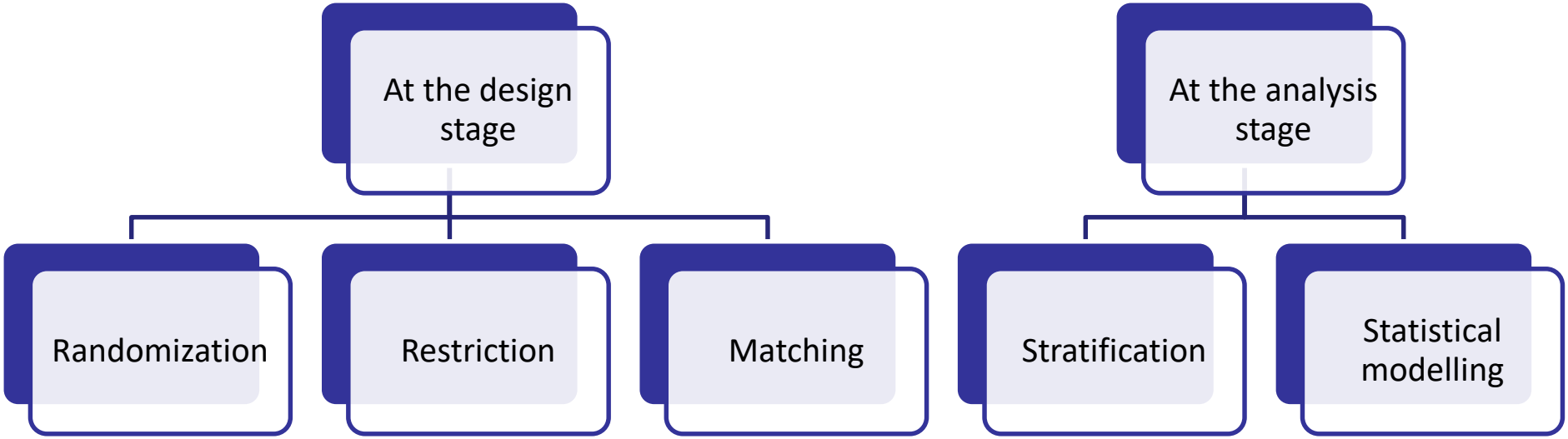


Table 2 fallacy

- Interpretation of coefficients in a regression model as if all representing the same type of effects.

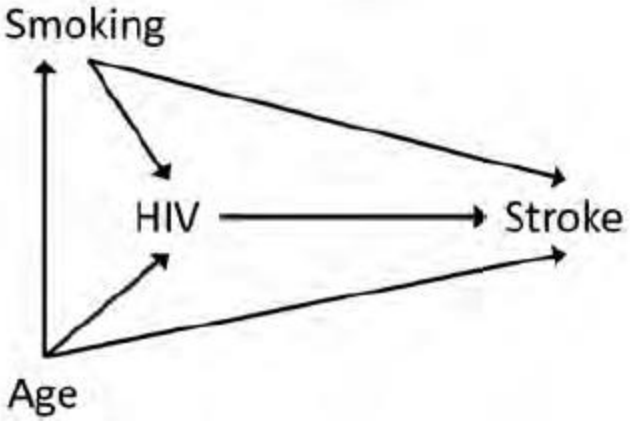


Figure 1. Causal diagram for the effect of human immunodeficiency virus (HIV) seroconversion on 10-year stroke risk, with confounding by smoking level and age.

$$\text{logit}(\text{Stroke}|\text{HIV}, \text{Smoking}, \text{Age}) = \beta_0 + \beta_1 \times \text{HIV} + \beta_2 \times \text{Smoking} + \beta_3 \times \text{Age}.$$

Westreich D, Greenland S. Am J Epidemiol. 2013 Feb 15;177(4):292-8. doi: 10.1093/aje/kws412.

Directed Acyclic Graphs: definition

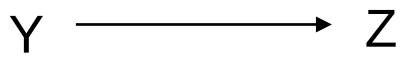
Directed	Acyclic	Graph
Variables connected with arrows	**Without** feedback loops	Abstract representation of an object and its relationships

*Non-parametric diagrammatic representations of the assumed data-generating process for a set of variables (and measurements thereof) in a specified context**

*Tennant PWG et al. Int J Epidemiol. 2021 May 17;50(2):620-632. doi: 10.1093/ije/dyaa213

What is required?

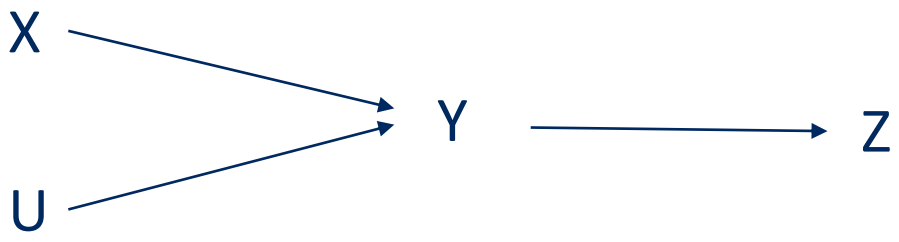
- Sufficient to represent the effect of Y on Z?



- Requires the inclusion of all common causes of the variables in the diagram. Otherwise it cannot be named a causal graph.

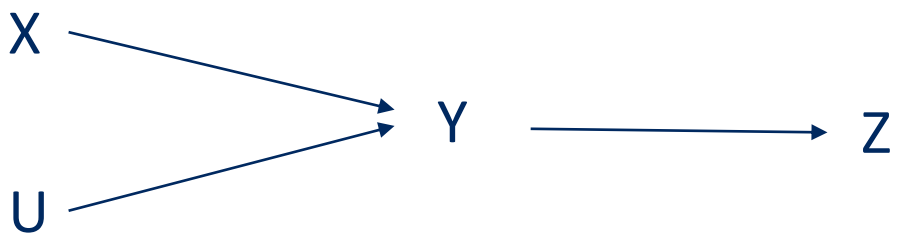
Main terms I

- Arrows are ***edges***;
- Variables are presented at the beginning and end of each arrow. They are ***nodes/vertices***;
- Variables connected by edges are ***neighbours/adjacents***;
- A ***path*** between X and Y is any given sequence of steps, starting in X and ending in Y, regardless of the arrows direction;
- A variable in the path between X and Y is referred to as *intercepting the path*.



Main terms II

- ***Descendents*** of variable X – variables affected by X, directly or indirectly.
- ***Sons*** of variable X – variables affected directly by X.
- ***Ascendents*** of variable X – variables affecting X, directly or indirectly.
- ***Parents*** of variable X – variables affecting X directly.



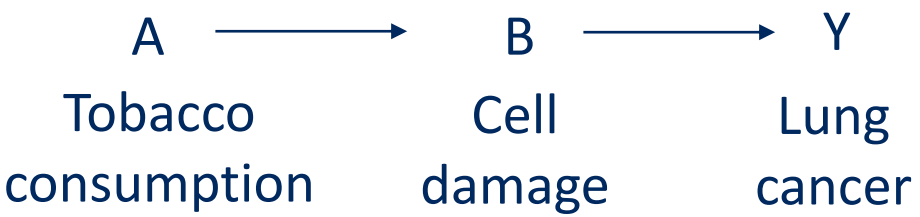
Causal models and statistical models

- Causal effects imply association;
- Absence of causal effects imply absence of association;

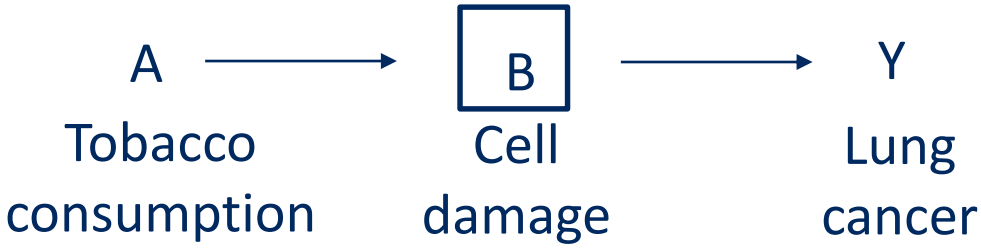
- Drawing a causal diagram implies drawing a statistical model.

Some rules – cause and effect

It is only possible to exclude an association between A and Y in the absence of an arrow of A to Y

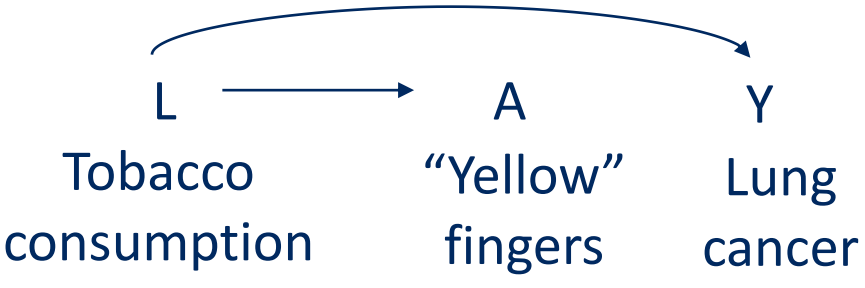


It is not required to represent the mediator to estimate the total effect of A in Y



The flux of the association between A and Y is interrupted when we condition on the mediator B

Some rules – common causes I

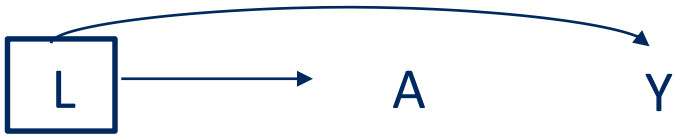


If we analyse data, will A and Y be associated?

It is not possible to exclude an association between A and Y when there is a common cause L, even if there is no arrow from A to Y

In other words, common causes might lead to **confounding**

Some rules – common causes II



Tobacco consumption

"Yellow" fingers

Lung cancer

What happens when we condition on L?

A and Y are **independent** when we **condition** on L

More broadly, the flux of the association from A to Y is **interrupted when we condition on the common cause L**

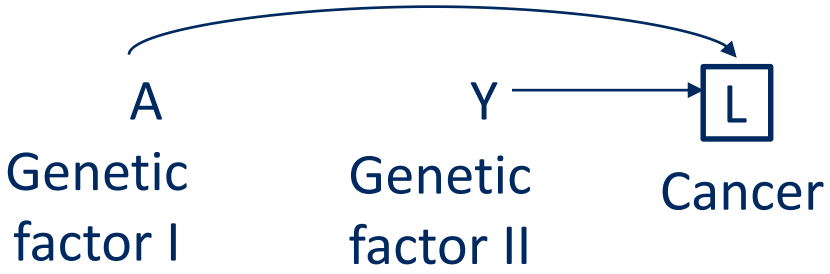
Some rules – common effects I

If we analyse data, will A and Y be associated?

L is a **collider**, a common effect of two variables.

Common effects do not create associations.

Some rules – common effects II



If we analyse data, will A and Y be associated?

It is not possible to exclude an association between A and Y when there is a common effect, L, and **we condition on that common effect**.

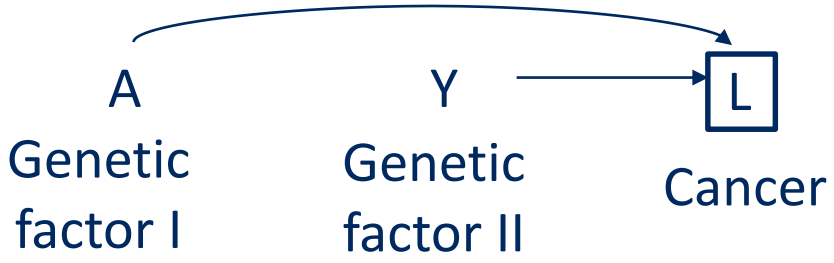
Selection bias corresponds to conditioning on a common effect

Common effects rules apply to colliders and their descendants.

Structural causes of association

Cause and effect

Common causes



Conditioning on a common effect

D-separation rules

1. If there are no variables being conditioned on, a path is blocked if and only if two arrowheads on the path collide at some variable on the path.

2. Any path that contains a non-collider that has been conditioned on is blocked.

3. A collider that has been conditioned on does not block a path.

4. A collider that has a descendant that has been conditioned on does not block a path.

D-separation and independence

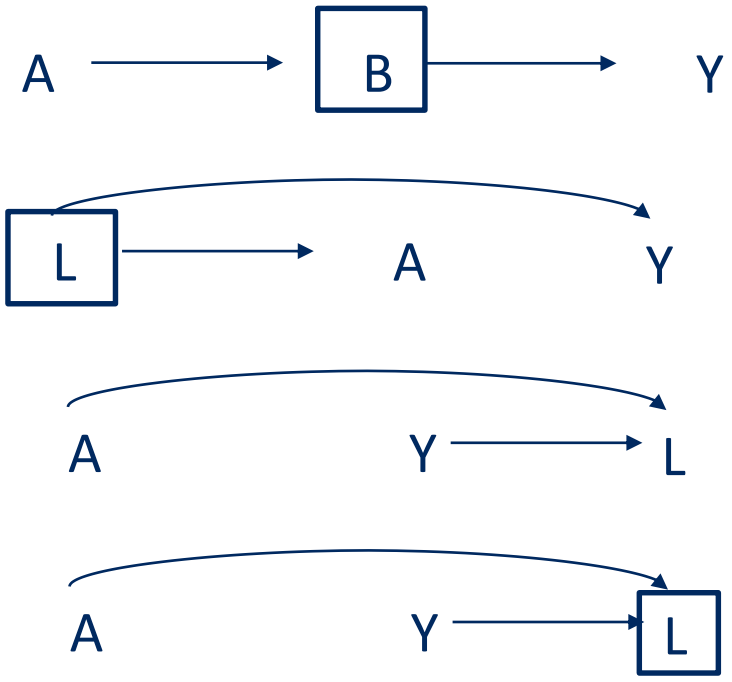
- Two variables are d-separated if all paths between them are blocked (otherwise they are d-connected).

- Two variables are marginally independent if they are d-separated without conditioning on other variables.

- Two variables are conditionally independent if they are d-separated after conditioning on other variables.

Questions

Is the path between A and Y open or blocked?



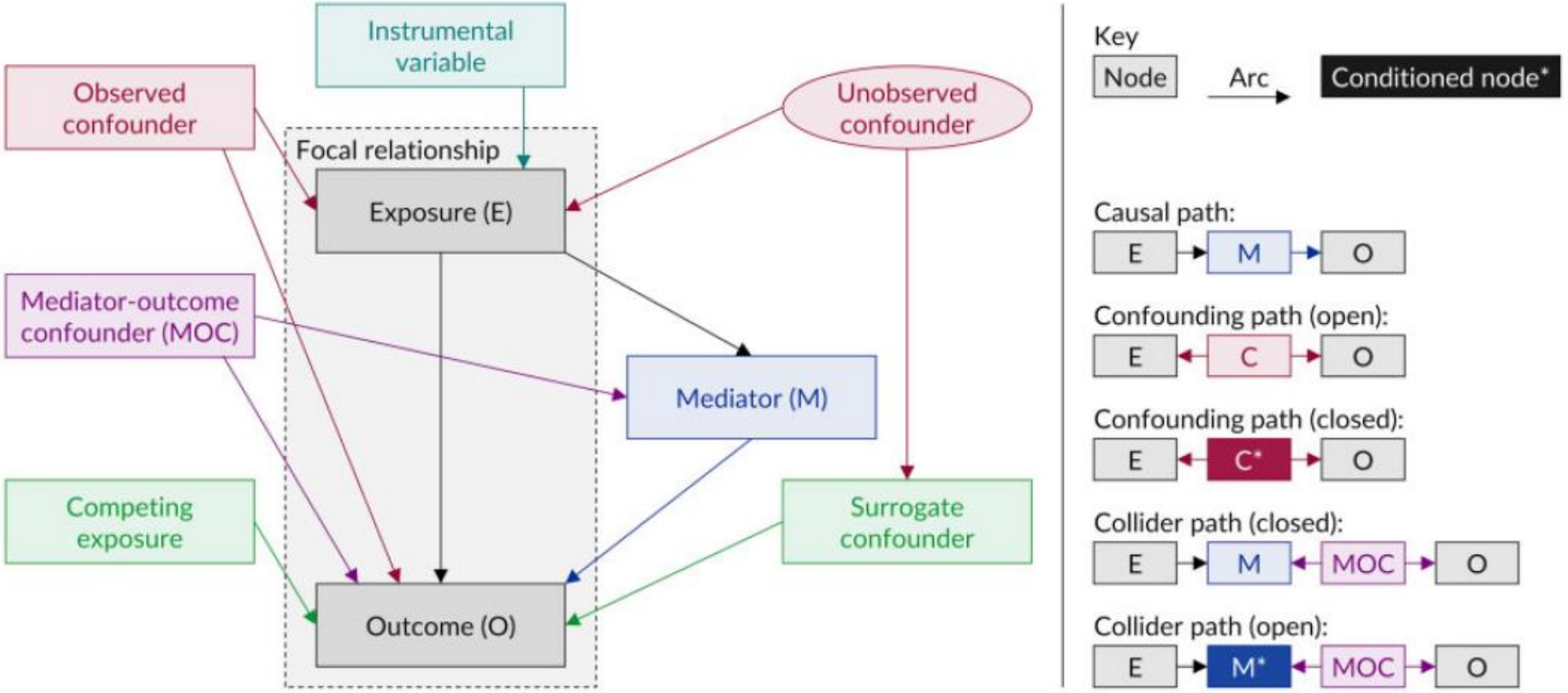
Blocked

Blocked

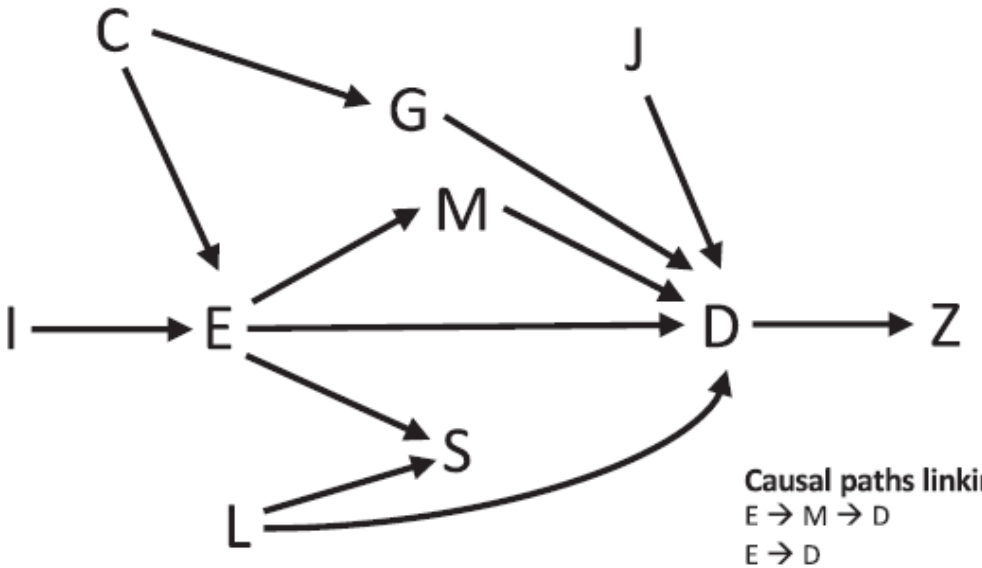
Blocked

Open

Main components



Tennant PWG et al. Int J Epidemiol. 2021 May 17;50(2):620-632. doi: 10.1093/ije/dyaa213



Causal paths linking E and D:

E → M → D

E → D

Non-causal paths linking E and D and how to block them:

E ← C → G → D (block by controlling for C or G)

E → S ← L → D (blocked provided we do not control for S)

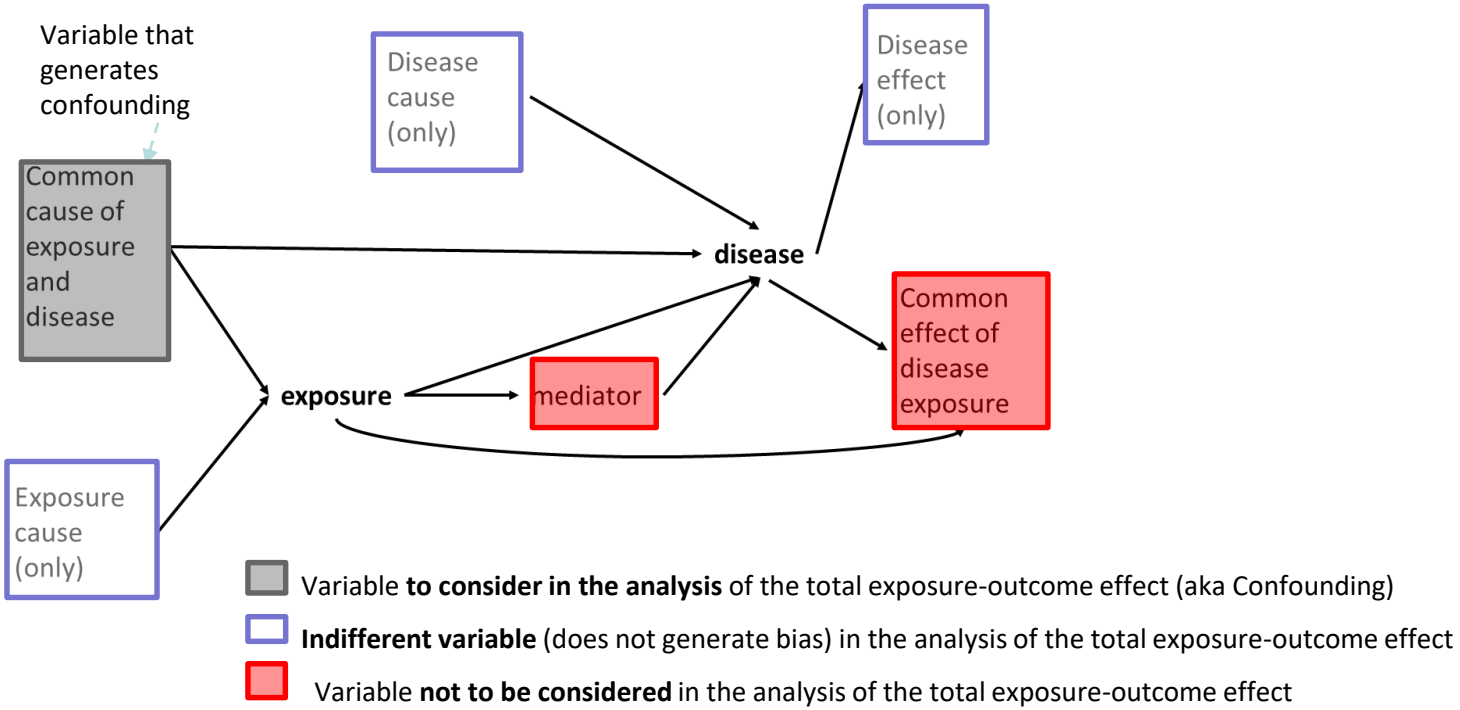
Key terms:

- C confounds the association of E and D.
- G can be controlled to block the confounding path between E and D.
- M partially mediates the effect of E on D.
- S is a collider on a non-causal path between E and L, and therefore a collider on a non-causal path between E and D. Controlling or restricting on S will create a biased association between E and D.
- Z is a descendant of D.
- I is an instrumental variable, such as randomization, for the effect of E on D.
- J causes D and will therefore be an effect modifier of any other cause of D on at least one scale (additive or multiplicative).

Digitale JC et al. J Clin Epidemiol. 2022 Feb;142:264-267. doi: 10.1016/j.jclinepi.2021.08.001.

Causal diagrams - overview

If we want to measure the **total effect** of exposure on disease there are variables that need to be included in the analysis and variables that should not be included (or conditioned on).



Slide from Baltazar Nunes

How to built a DAG I

Table 1. Summary of ESC-DAGs protocol

Stage	Purpose	Process
Mapping	To apply graph theory to the conclusions of each study. This creates an 'implied graph' (IG) which acts as a transparent structural template for translation into a DAG.	1. Outcome variable of interest is set as DAG outcome(s). 2. Exposure variable(s) of interest is set as DAG exposure(s). 3. A directed edge is drawn originating from the exposure(s), terminating at the outcome(s). 4. All control variables are entered as unassigned variables. 5. A directed edge is drawn originating from each control to the exposure(s) and outcome(s). 6. Mediators, instrumental variables etc. are mapped as per the study's conclusions. 7. The IG is saturated by drawing directed or undirected edges between all confounders (direction does not matter until the translation stage). The recombination process can be performed at this stage to help simplify an overly complex IG.
Translation	To apply causal theory to each relationship in the IG. This creates the DAG for the study. Each relationship in the IG is assessed under sequential causal criteria and a counterfactual thought experiment (See causal criteria sections in the text for detailed discussion).	The posited relationship and its reverse are both assessed. Edges may be retained as posited, reversed, or as bi-directional. If not, they are deleted. All retained edges are entered into the directed edge index. 1. Temporality—does the posited cause precede effect? (If 'yes', proceed to next criterion. If not, assess reverse relationship.) 2. Face-validity—is the posited relationship plausible? (If 'yes', proceed to next criterion. If not, assess reverse relationship.) 3. Recourse to theory—is the posited relationship supported by theory? (Always proceed to the counterfactual thought experiment.) 4. Counterfactual thought experiment—is the posited relationship supported by a systematic thought experiment informed by the POF? (Once completed, always assess the reverse relationship unless already assessed.)

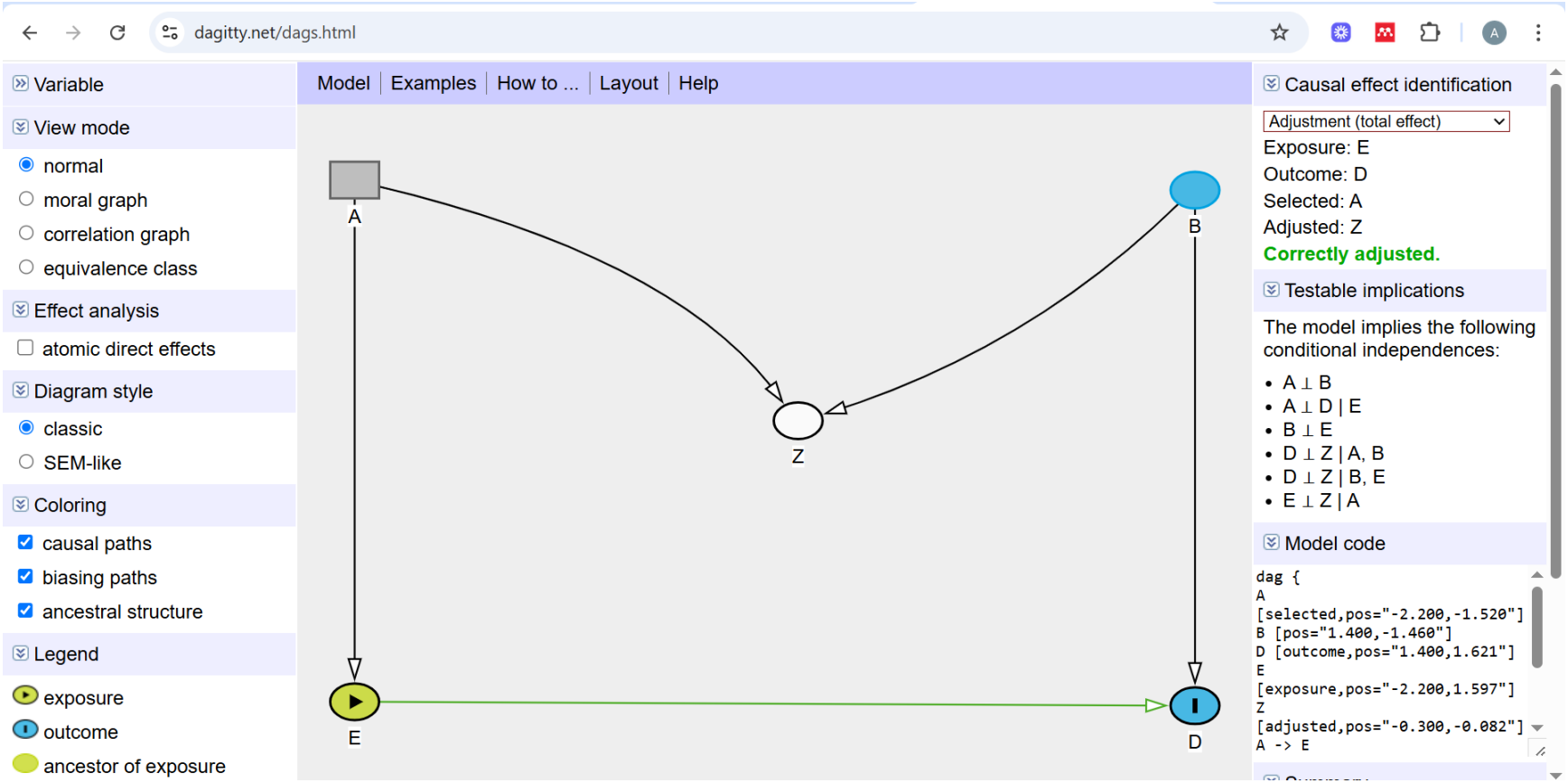
Fersugon et al. Int J Epidemiol 2020; 49(1): 322-29. doi: 10.1093/ije/dyz150

How to build a DAG II

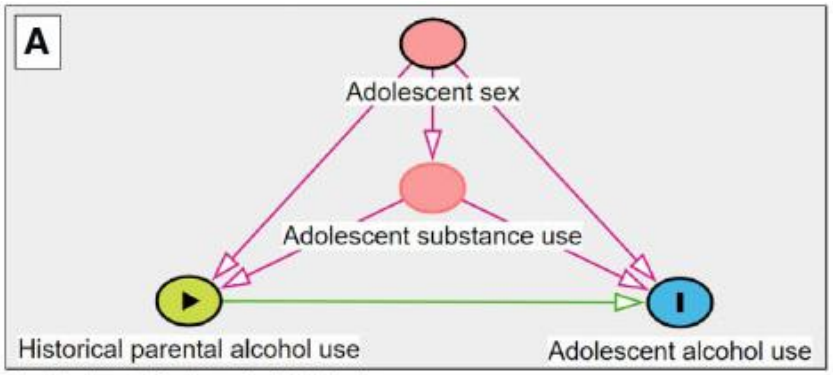
Integration 1: synthesis	To combine the translated DAGs into one by synthesising all indexed directed edges.	1. A new DAG is created to serve as the integrated DAG (I-DAG). 2. The focal relationship is added to the I-DAG (as per mapping steps 1–3). 3. Each indexed directed edge pertaining to the focal relationship (including its corresponding node) is added to the diagram. 4. Each indexed directed edge pertaining to other nodes is added (e.g. between confounders). 5. Conceptually similar nodes should be grouped together in virtual space to aid the recombination process.
Integration 2: recombination	To combine nodes for either practical reasons (i.e. to reduce complexity) or substantive reasons (i.e. to establish consistency).	1. Is there theoretical support for combining two variables/nodes? 2. Do the conceptually related nodes have similar inputs and outputs (i.e. do they 'send to' and 'receive from' the same nodes)?

Ferguson et al. Int J Epidemiol 2020; 49(1): 322-29. doi: 10.1093/ije/dyz150

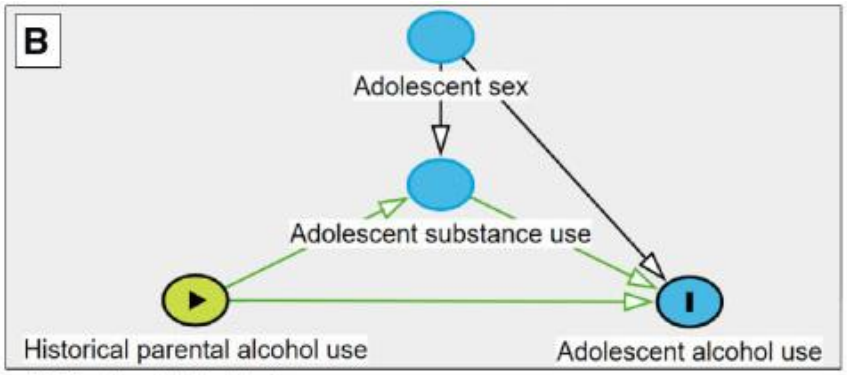
How to build a DAG III



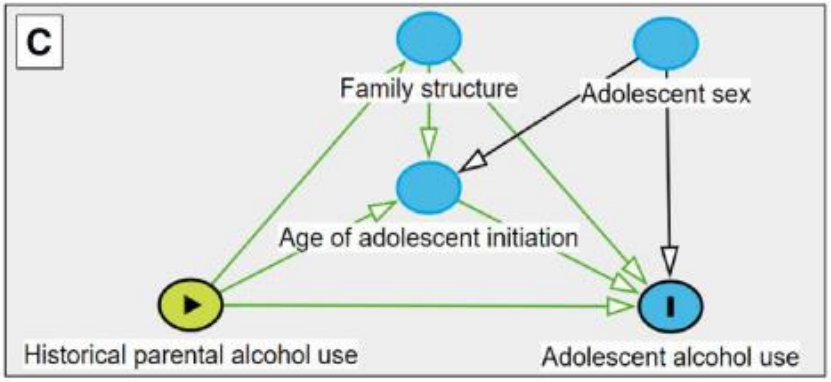
DAG – Example I



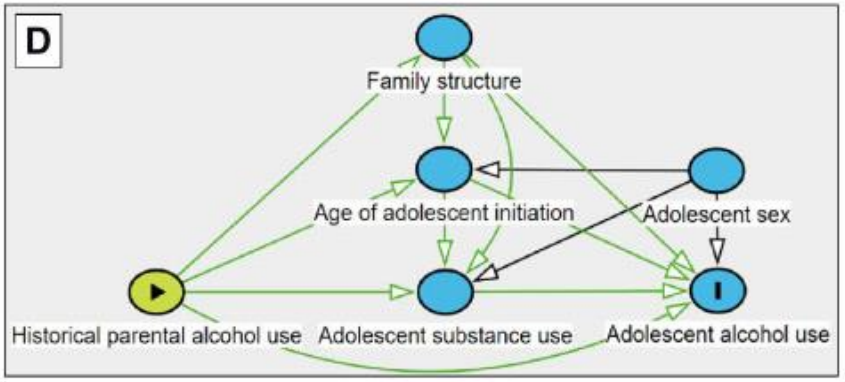
Implied graph for hypothetical study



DAG for hypothetical study



DAG for Seljamo et al study



Integrated DAG for hypothetical and Seljamo et al studies

Ferguson et al. Int J Epidemiol 2020; 49(1): 322-29. doi: 10.1093/ije/dyz150

Other applications and extensions

- Selection bias;
- Measurement error and information bias;
- Effect modification;
- Other types of causal diagrams - *Single World Intervention Graphs* (SWIGs) – explicitly connect the potential outcome framework with DAGs.

Use of DAGs

Table 1. Summary information regarding the reporting of estimands and adjustment sets in the 234 included studies, and regarding the reporting and features of the largest DAG in the 144 studies with ≥1 DAG

DAG reporting and features[a]	*n*	% (*n* = 144)
DAG available	144	100%
Single DAG available	116	81%
Multiple DAGs available	28	19%
DAG includes one or more unobserved variables	53	37%
DAG includes[a] one or more specific unobserved variables	27	19%
DAG includes[a] one or more generic unobserved variables	29	20%
Visually arranged so all arcs flow in the same direction	49	34%
Top-to-bottom	5	3%
Left-to-right	22	15%
Corner-to-corner	22	15%
Authors provide citations for one or more arcs	8	6%

DAG nodes and arcs	Median	IQR	Range
Number of nodes	12	9–16	3–28
Number of arcs	29	19–42	3–99
Ratio of arcs-to-nodes	2.3	1.8–3.0	1.0–5.8
Saturation (%)[b]	46	31–67	12–100

Reporting of estimand(s) and adjustment set(s)	*n*	% (*n* = 234)
Report one or more estimand(s) of interest	48	21%
Report seeking total causal effects	18	8%
Report seeking direct causal effects	12	5%
Report seeking multiple effects	18	8%
Report DAG-implied adjustment set(s)	115	49%
Report results of DAG-implied adjustment set(s)	101	43%
Report as primary results	95	41%
Report results of other or unclear adjustment set(s)	171	73%
Report as primary results	159	68%
Use additional statistical criteria for variable selection	42	18%

[a]Details are for the largest DAG reported in each study.
[b]The saturation percentage represents the proportion of all possible arcs that have been included.

Tennant PWG et al. Int J Epidemiol. 2021 May 17;50(2):620-632. doi: 10.1093/ije/dyaa213

Use of DAGs: recommendations

- The **focal relationship(s) and estimand(s)** of interest should be stated in the study aims;
- The DAG(s) for each focal relationship and estimand of interest should be **available**;
- DAGs should include all relevant variables, including those where direct **measurements are unavailable**;
- Variables **should be visually arranged** so that all constituent arcs flow in the same direction;
- **Arcs should generally be assumed** to exist between any two variables;
- The **DAG-implied adjustment set(s)** for the estimand(s) of interest should be clearly stated;
- The **estimate(s) obtained from using the unmodified DAG-implied** adjustment set(s) should be reported;
- **Alternative adjustment** set(s) should be justified and their estimate(s) reported separately.

Tennant PWG et al. Int J Epidemiol. 2021 May 17;50(2):620-632. doi: 10.1093/ije/dyaa213

Overview

→ Easy, simple to learn tool;

→ Guide study design and analysis;

→ Supports recognition of bias;

→ Enable clear communication.

→ Limitations in prior knowledge;

→ No information on magnitude and form;

→ Complex whith time dependencies;

→ Primarly applied to causal questions.

Take home messages

- DAGs provide a simple and transparent representation about the assumed causal relationships between variables;
- They are a powerful tool in causal inference and are best used in conjunction with other approaches in the field;
- We can use DAG to identify and address sources of biases;
- DAGs rely on existing knowledge, they are praised for their transparency not their perfection.

Selected references

- Digitale JC, Martin JN, Glymour MM. Tutorial on directed acyclic graphs. J Clin Epidemiol. 2022 Feb;142:264-267. doi: 10.1016/j.jclinepi.2021.08.001.
- Ferguson K et al. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. Int J Epidemiol 2020; 49(1): 322-29. doi: 10.1093/ije/dyz150.
- Glymour MM and Greenland S. Causal diagrams. In Modern Epidemiology, 3rd edition, Rothman KJ, Greenland S, and Lash T, eds. Lippincott-Raven (Book chapter, 2008).
- Hernan M. Causal Diagrams: Draw Your Assumptions Before Your Conclusions. Available from: https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your [accessed 2025-07-09].
- Hernan M, Robins J. Causal inference: What if. Boca Raton: Chapman & Hall/CRC, 2023. Available from: https://miguelhernan.org/whatifbook [accessed 2025-07-09].
- Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench. 2012 Spring;5(2):79-83.
- Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, Harrison WJ, Keeble C, Ranker LR, Textor J, Tomova GD, Gilthorpe MS, Ellison GTH. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol. 2021 May 17;50(2):620-632. doi: 10.1093/ije/dyaa213.
- Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol. 2013 Feb 15;177(4):292-8. doi: 10.1093/aje/kws412.

7th Statistics on Health Decision Making: Epidemiology

July 10 › 11, 2025 | University of Aveiro

DIRECTED ACYCLIC GRAPHS AS A CAUSAL INFERENCE TOOL:

From principles to applications

Andreia Leite

National Institute of Health Research Doctor Ricardo Jorge

NOVA National School of Public Health